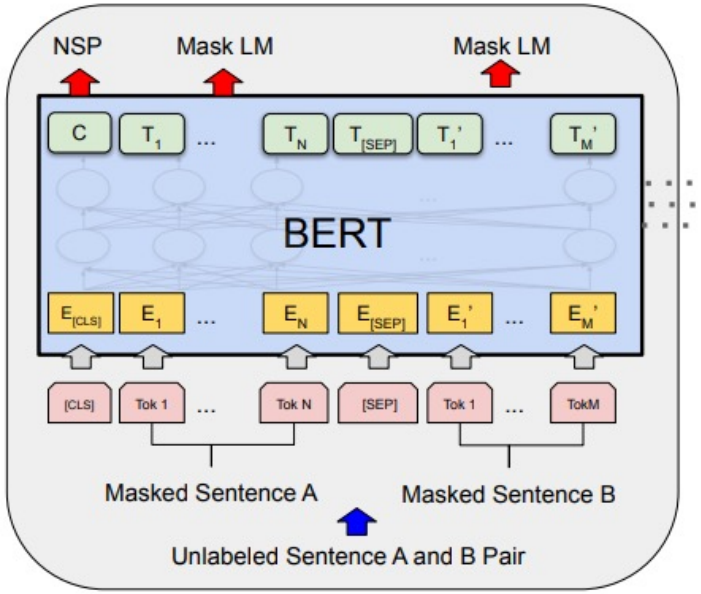
Generative Video Transformer: Can Objects be the Words?

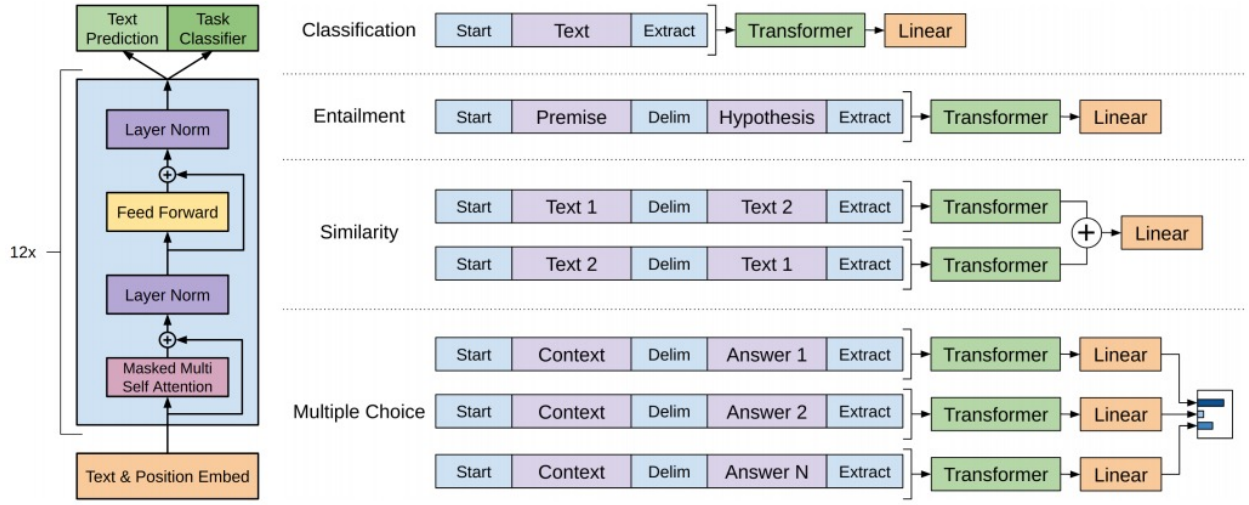
Yi-Fu Wu[1], Jaesik Yoon[1,2], Sungjin Ahn[1]

[1]Rutgers University, [2]SAP

Motivation



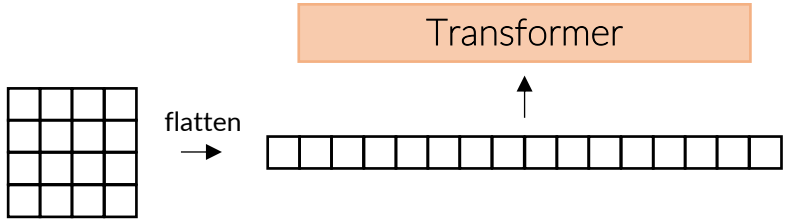
BERT (Devlin et al., 2019)

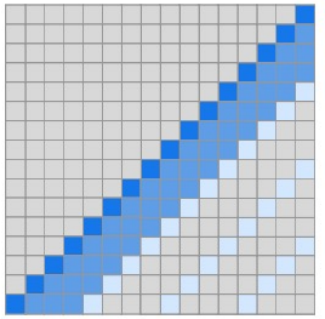


GPT (Radford et al., 2018)

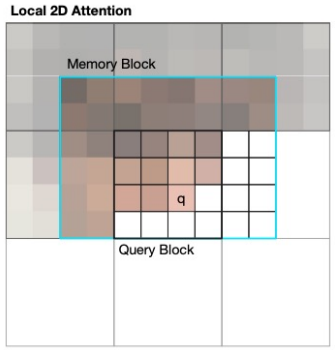
- Can a transformer-based architecture be effective for generative pre-training of visual scenes for video generation and understanding?

Previous Work

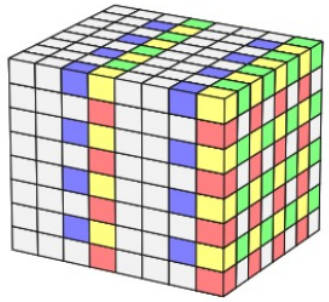




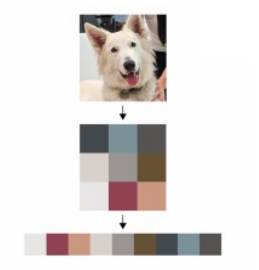
Child et al., 2019



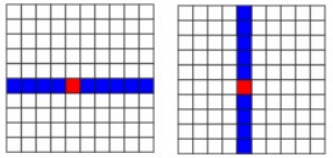
Parmer et al., 2018



Parmer et al., 2018



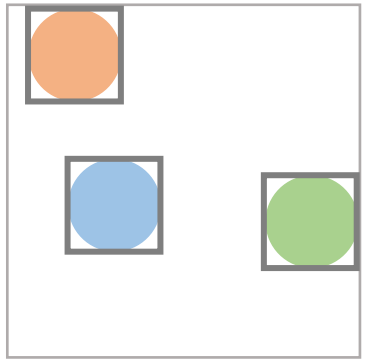
Chen et al., 2020



Ho et al., 2019

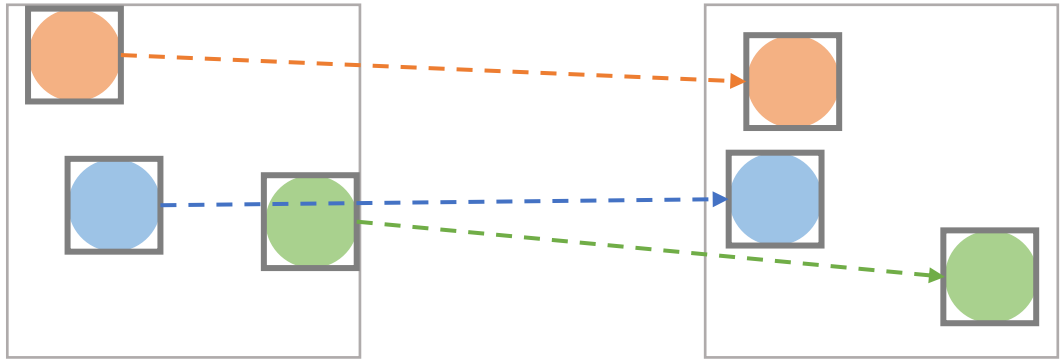
Dosovitskiy et al., 2021

Our Approach



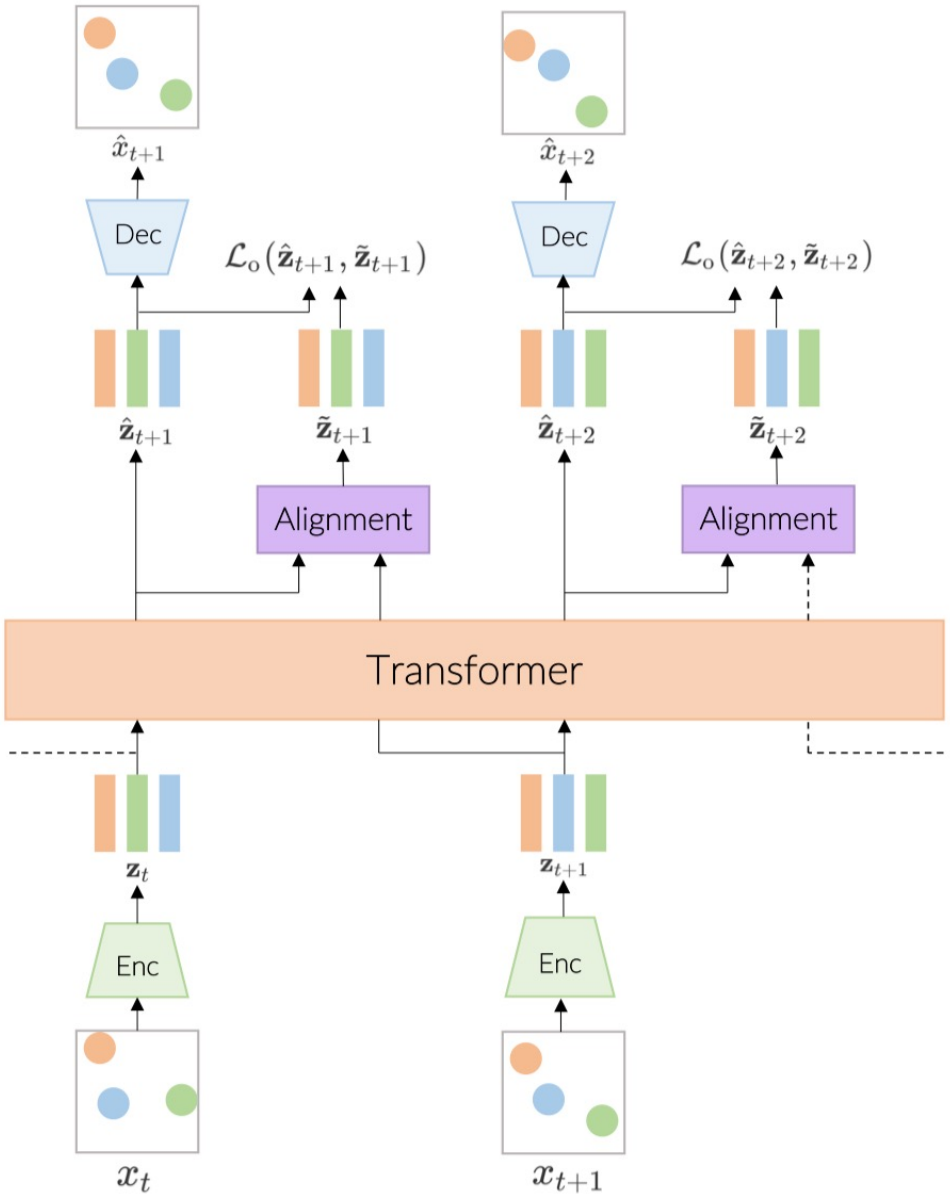
- Tokenize an image into its consitutent object representations and use these representations as input to the transformer

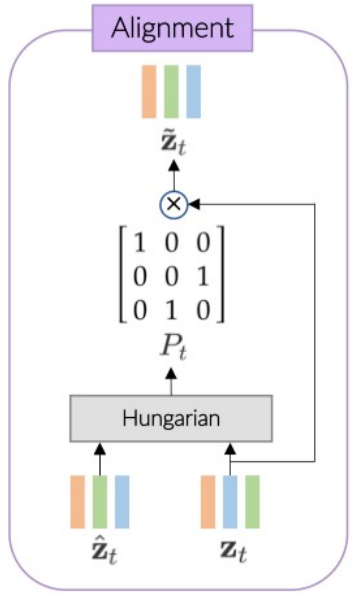
Our Approach

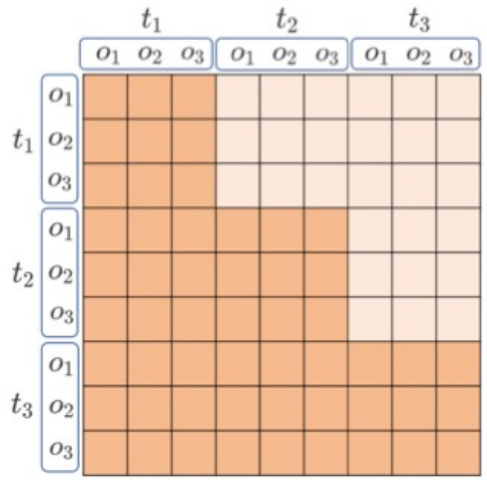


- Predict the entire image at once by generating all the objects in an image simultaneously given their previous states
- Align objects between frames based on object location, leveraging SPACE (*Lin et al., 2020*) an unsupervised object representation model that outputs explicit bounding box information

Object-Centric Video Transformer



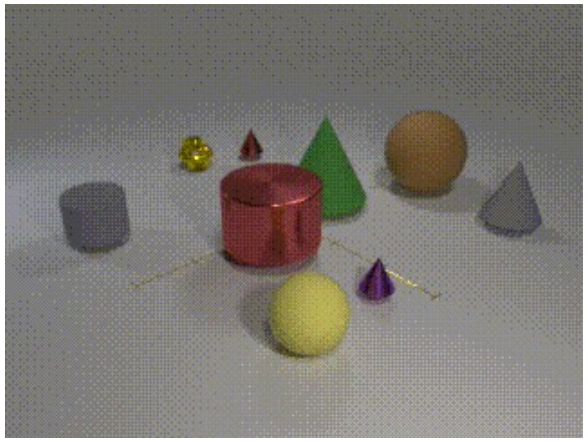


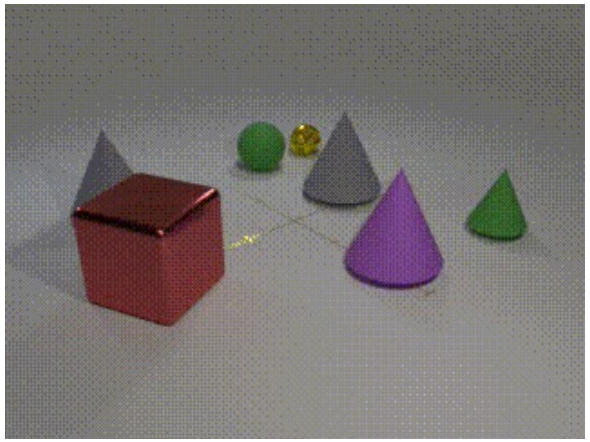


Experiments - Datasets

- Bouncing balls

- CATER

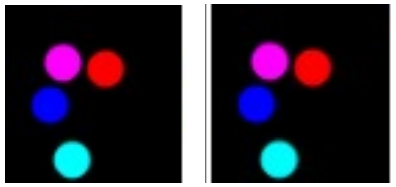




Experiments – Bouncing Ball

Table 1. Average next-step prediction color change accuracy

	MOD1	MOD2	MOD3	MOD1234
GSWM	71.69	17.51	14.63	11.72
LSTM+GNN	73.64	69.08	22.30	51.38
SVVT	37.53	18.23	11.96	29.47
CONVVT	88.31	82.83	46.49	67.29
CONVVT-AR	8.70	4.20	3.25	6.10
OCVT-AR	78.70	76.99	54.49	64.97
OCVT (OURS)	**89.61**	**88.18**	**82.70**	**78.43**



Ground Truth Generation

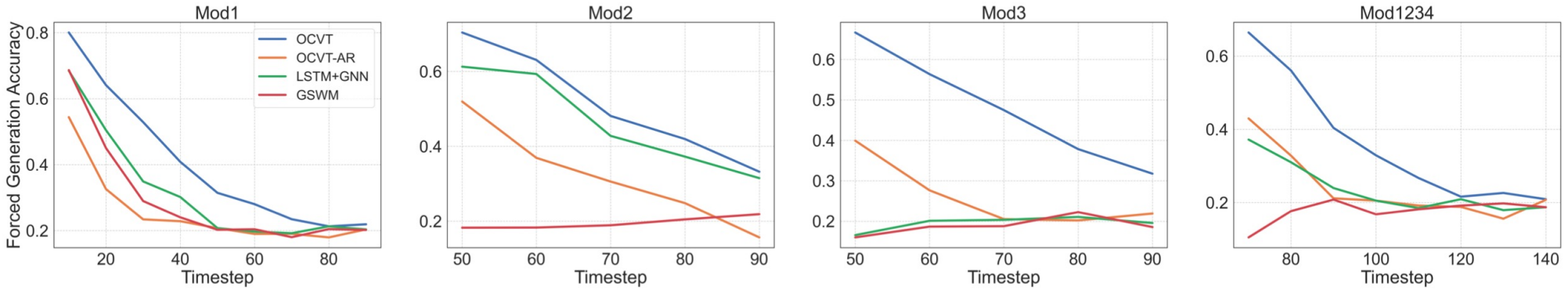


Figure 5. Forced Generation Accuracy

Experiments - CATER

Table 2. CATER results

	Top 1 ↑	Top 5 ↑	L1 ↓
Ding et al	70.6	93.0	0.53
Ding et al w/ L1	74.0	94.0	0.44
Hopper	73.2	93.8	0.85
OPNet	74.8	-	0.54
OCVT (ours)	**76.0**	94.4	0.45
OCVT w/ L1 (ours)	75.9	**95.3**	**0.39**

Conclusion

- OCVT is able to generate future frames of videos with complex long-term dependencies
- Learned representations are useful for downstream tasks

- *Please refer to our paper for more details*